Bachelor's Thesis

Submitted in Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Arts in Computational Linguistics

Alignment Weighting for Short Answer Assessment

Author:
Björn Rudzewitz

Supervisor:
Prof. Dr. Detmar Meurers

Seminar für Sprachwissenschaft
Eberhard-Karls-Universität Tübingen

August 2015

I hereby declare that this paper is the result of my own independent scholarly work. I have acknowledged all the other authors' ideas and referenced direct quotations from their work (in the form of books, articles, essays, dissertations, and on the internet). No material other than that listed has been used.

Tübingen, August 31, 2015

Björn Rudzewitz

Contents

Abstract

This work describes the extension of a short answer assessment system with novel alignment weighting features. Two conceptually different weighting schemes are proposed: a general linguistic and a task-specific weighting. The general linguistic weighting is exemplified by the implementation of part of speech features that indicate the quantity of syntactic classes of aligned elements. The task-specific weighting scheme is operationalized by implementing question type features for the question a short answer was given to. Furthermore, a hybrid approach is explored with the implementation of term weighting features trained on the task context material.

The effectiveness of the new features and their interactions are experimentally tested on a corpus of learner data produced by learners of German at the university level. Statistical significance tests show that the new features improve the system performance significantly up to 90.0% accuracy for the task of the binary classification of the correctness of learner answers to reading comprehension questions. An exemplary analysis of the question type specific system accuracy shows that the new features help the system overcome previously reported problems for question forms that allow a high variation in the surface realization of the answer. A comparison with related work reveals that the performance of the system augmented with the new features is highly competitive with state-of-the-art approaches towards short answer assessment.

List of Figures

List of Tables

1 Introduction

Reading comprehension constitutes a common task in real-life foreign language learning. The language learners are asked to read a text in the foreign language and provide answers in free-text form to questions asking about the content of the reading text. This requires the learners to on the one hand read and understand the content of the text, and on the other hand to understand the concrete task expressed as a question in order to appropriately fulfill the requirements of the task. The process of forming the answer to a question requires the learner to extract relevant information from the text and to manipulate the form of this content by applying linguistic knowledge about the foreign target language.

This process is prone to a range of errors. The learners may not sufficiently understand the reading text. Therefore, they may use irrelevant material for answering the question. Another source of error can be the misunderstanding of a question to the text and thus a misconception of the task. This can have the same consequences as misunderstanding the reading text. Even if a learner understands the reading comprehension text and the task, errors or mistakes may arise from a lack of competence or performance concerning linguistic structures in the foreign language. Learners may over- or underuse certain linguistic rules, or they may simply copy passages from the text without adapting this textual material. The result of these processes is a spectrum of learner responses to reading comprehension. This spectrum has multiple related dimensions: one dimension is concerned with the correctness of the answer. The question has to be asked what criteria distinguish a correct answer from an incorrect answer. A related dimension displays the deviation of the learner language from the standard language as produced by native speakers of this language. The learner language may represent an intermediate state of language competence in the target language. From the perspective of a native speaker, the learner language may exhibit form errors (for example spelling mistakes) or grammar errors (for example a wrong tense form).

While human language teachers are capable of dealing with this form of language and can infer the (in)correctness of an ill-formed learner response, the question arises whether and to what extent machines can perform this task of content assessment. The research area concerned with this task is the field of Short Answer Assessment (SAA, (Ziai, Ott, & Meurers, 2012), (Burrows, Gurevych, & Stein, 2015)). The present work is part of a specific field of SAA concerned with educational assessment in the context of foreign language learning. In this context two problems have to be addressed at the same time: on the one hand it is necessary to implement a content assessment approach, and on the other hand it is necessary to robustly deal with the variation in learner language. This complex task only becomes feasible if a concrete task context is given. However, few work ((S. Bailey & Meurers, 2008), (Meurers, Ziai, Ott, & Bailey, 2011), (Ziai & Meurers, 2014), (Horbach, Palmer, & Pinkal, 2013)) has been done on actually using the task context in short answer assessment. This work takes a step into this direction and proposes new ways of making use of the concrete task context.

The basic approach towards the problem of short answer assessment is to compare student answers to manually defined target answers, and to determine the degree of similarity between them on different levels of abstraction. This also forms the basis of the approach in the present work. Given a reading text with questions and learner

and target answers to this question, the automatic short answer assessment system CoMiC-DE (Meurers, Ziai, Ott, & Kopp, 2011) aligns elements found to be similar on different levels of linguistic abstraction. The research question to be explored in the present work is to what extent it is possible to weight these alignments by on the one hand using general properties of the foreign language, and by on the other hand incorporating the specific task context in which a learner answer was produced. To show the impact of such weightings in practice, certain representative features encoding these properties are implemented. For the general linguistic weighting, the implementation of part of speech features is conducted. As a hybrid approach between general linguistic weighting and task-specific weighting, a term-frequency inverse document frequency (*tf.idf*) weighting is presented. While this measure is applicable in every context of a language, its practical realization in this work is task-specific since it encodes information of the reading comprehension texts and thus the task context. Another weighting is proposed as a purely task-specific weighting method: question type features are implemented to shallowly encode information about the concrete task setting.
In order to test the effect of these weighting methods on the system performance, certain components are needed that will be discussed in the following sections. Section 2 presents a corpus of reading comprehension tasks collected from language learners of German. In section 3 the baseline short answer assessment system used as a testbed for the implementation of the new alignment weighting features is discussed. Section 4 presents the new general linguistic alignment weighting features before section 5 discusses the task-specific weighting features. Section 6 shows experimental results of the performance of the baseline system when augmented with the new features. These results are put into a wider context in section 7 before section 8 summarizes the main findings and draws routes for future work in this area.

2 Data

The experiments reported in this work data are based on the Corpus of Reading Comprehension Exercises in German (CREG) ((Meurers, Ott, Ziai, et al., 2010), (Ott, Ziai, & Meurers, 2012), (Meurers, Ziai, Ott, & Kopp, 2011)). CREG represents a collection of reading comprehension exercises in German consisting of reading texts, reading comprehension questions, student answers, target answers, and meta data.
The data was collected via the web-based tool WELCOME (Meurers et al., 2010) at the German programs at the Kansas University (KU) and The Ohio State University (OSU). The corpus is a learner language corpus that represents the language abilities of learners of German at different levels of proficiency. The result is that the learner language contains a considerable amount of form errors (see (Meurers, Ziai, Ott, & Kopp, 2011)) that need to be dealt with by the content assessment system. Detailed information about the individual learners is encoded in the corpus via meta data, allowing a longitudinal perspective on the language proficiency development. For a detailed itemization of these properties, refer to (Meurers et al., 2010).
A main consideration of the corpus design is the explicit encoding of the task context for the reading comprehension exercises (Meurers, Ziai, Ott, & Kopp, 2011). For this purpose, the corpus not only contains the student and target answers, but also the reading comprehension text used as a textual basis for the task, along with meta data about all of the named components. According to (Meurers et al., 2010), reading

comprehension questions are included in the CREG corpus only if all the knowledge needed to answer this question is encoded explicitly in the corresponding text. This limits the space of possible productions by learners in that world knowledge should not influence the answering of a question. Furthermore, the encoding of the reading comprehension questions themselves already is a partial representation of the task context. Section 5 presents previous work of a study highlighting the importance of taking into account the different question types as an approximation of the task context. Motivated by these insights, section 6.2 reports on the effect of the performance of a short answer assessment system when augmented with features representing a certain question type.

Among the 1,517 questions, the corpus contains 36,335 student answers to these questions (Ziai & Meurers, 2014). Each student answer is labeled with a binary diagnosis indicating its semantic (in)correctness independent of the surface realization given the explicit task context. The diagnosis was assigned by two independent annotators to every answer. (Ott et al., 2012) reported substantial agreement for the annotation of binary labels. In addition to the binary diagnosis, each student answer was assigned a detailed label that specifies the potential type of semantic error that was made in this answer (see (Meurers et al., 2010)). The binary correctness feature can take the values *correct* or *incorrect*, and the detailed label can take the values *correct answer*, *missing concept*, *extra concept*, *blend* (missing concept and extra concept), or *non-answer*. This answer assessment taxonomy is based on the taxonomy proposed by (S. Bailey & Meurers, 2008), the only difference being that the *alternate answer* label was not allowed as a feature value for the detailed diagnosis feature. In (S. Bailey & Meurers, 2008), this label was used for a semantically plausible answer not represented in the target answers. For the CREG corpus, the teachers grading the student answers instead were asked to encode additional target answers (Meurers et al., 2010). For this work, however, only the binary diagnosis is used.

As a reference, the corpus contains 2,057 target answers specified by the above mentioned annotators (Ziai & Meurers, 2014). The majority of the student answers is associated with a target answer to which it is compared by the content assessment system. The target answers represent spelled-out sample solutions to the reading comprehension questions.

The present study uses subsets of the CREG corpus for the experimental testing. Each data subset consists of two parts: the KU part (collected a the Kansas University), and the OSU part (collected at The Ohio State University). In order to make the results comparable to previous work ((Meurers, Ziai, Ott, & Kopp, 2011), (Ziai & Meurers, 2014), (Hahn & Meurers, 2012), (Horbach et al., 2013)), the CREG-1032 corpus is used. This subset of CREG can be seen as a representative sample of the corpus since it is both balanced in terms of the binary diagnosis and at the same time both human annotators agreed on the diagnosis label. Furthermore, the OSU and KU subsets are also balanced themselves. This portion of CREG is the only data set for which results of other systems are available. Section 7 compares the results of the present work to previous work on CREG-1032. Another data set used for the experiments in the present study is the CREG-5K data set. It contains 5138 student answers to 877 questions. CREG-1032 is a subset of CREG-5K. CREG-3620 is another subset of CREG-5K where all questions with a surface form appearing in the corresponding CREG-1032 data set and all the answers to these questions were removed. This data set was used in order to have a data set with questions that are independent

of the questions in CREG-1032, which were used to train the question type detector. The numbers in the table indicate the raw frequencies of the elements. Especially for CREG-1032 the number of target answers is higher than reported in previous publications (Meurers, Ziai, Ott, & Kopp, 2011), where they reported 136 (KU) and 87 (OSU) target answers. These lower numbers however only reflect the actually selected target answers and not all the target answers available in the CREG-1032 corpus. The new system uses the edit distance (Levenshtein, 1966) to select a target answer for cases where the target answer was not annotated for a student answer. In this process it iterates over all relevant target answers to select the closest. For this reason, the total number of target answers is reported here.

data set	**# questions**	**# student answers**	**# target answers**
CREG-1032-KU	117	610	180
CREG-1032-OSU	60	422	147
CREG-3620-KU	89	735	181
CREG-3620-OSU	585	2885	705
CREG-5K-KU	214	1814	382
CREG-5K-OSU	663	3324	875

Table 2.1: Data distribution of CREG subsets used in the present study.

3 Baseline System

This work is conducted on the basis of the CoMiC-DE system (Meurers, Ziai, Ott, & Kopp, 2011). CoMiC-DE is an adaption of CoMiC-EN (Meurers, Ziai, Ott, & Bailey, 2011) for German. CoMiC-EN in turn is a re-implementation of the CAM system ((S. M. Bailey, 2008), (S. Bailey & Meurers, 2008)).
These short answer assessment systems assess the meaning of a student answer to a reading comprehension question by establishing alignments on different linguistic levels of abstraction in order to use the quantity of alignments of different qualities to conduct a diagnosis in terms of semantic correctness of the student answer.
CoMiC-DE follows the same conceptual three-stage approach as its predecessor systems (see (S. Bailey & Meurers, 2008),(Meurers, Ziai, Ott, & Bailey, 2011),(Meurers, Ziai, Ott, & Kopp, 2011)). The system first enriches the input text with linguistic annotation before using the annotations to establish alignments between the student and the target answer. The quantity of different alignment types is then used to conduct a diagnosis of the student answer in a machine learning component. In the following, the different steps will be explained in detail since they form the basis for the present work.
The first step is the annotation of the textual material with linguistic information. Table 3.1 taken from (Meurers, Ziai, Ott, & Kopp, 2011) lists the NLP tools used for the different linguistic annotation tasks. Even though the tools are the same as reported in (Meurers, Ziai, Ott, & Kopp, 2011), the present work partly makes use of newer versions of them including newer models. This work uses MaltParser 1.8.1 (Nivre et al., 2007) instead of MaltParser 1.4. Also the MaltParser model was trained on TüBa-D/Z 9.0 (Telljohann, Hinrichs, Kübler, Zinsmeister, & Beck, 2012) instead of TüBa-D/Z 5.0 as in (Meurers, Ziai, Ott, & Kopp, 2011). Furthermore the present work makes use of OpenNLP 1.5.3 (Baldridge, 2005) instead of 1.4. Again the pre-trained

models provided by OpenNLP are used[1] (Baldridge, 2005).

Task	**NLP Tool**
Sentence Detection	OpenNLP[2] (Baldridge, 2005)
Tokenization	OpenNLP (Baldridge, 2005)
Lemmatization	TreeTagger (Schmid, 1994)
Spell Checking	Edit distance (Levenshtein, 1966), igerman98 word list
Part of Speech Tagging	(Schmid, 1994)
Noun Phrase Chunking	OpenNLP (Baldridge, 2005)
Lexical Relations	GermaNet (Hamp, Feldweg, et al., 1997)
Similarity Score	PMI-IR (Turney, 2001)
Dependency Relations	MaltParser (Nivre et al., 2007)

Table 3.1: NLP tools used in the CoMiC-DE system.

The system is implemented in the UIMA framework (Ferrucci & Lally, 2004). This architecture enables storing the information provided by the tools in table 3.1 in a multi-layer standoff format (Meurers, Ziai, Ott, & Bailey, 2011). Annotations from different layers can be combined with each other (as for part of speech tags annotated on tokens), or they can be independent (as for noun phrase chunks alongside sentences). One difference to the original CoMiC-DE system (Meurers, Ziai, Ott, & Kopp, 2011) in this point is that the new system makes use of uimaFit (Ogren & Bethard, 2009), which handles most of the XML components of UIMA automatically via a Java API. The linguistic annotation is performed for each student answer, corresponding target answer, and question.

After the annotation phase, the alignment phase is the next step in the system architecture. In this phase, the annotations created in step 1 are used to determine the similarity of student and target answers on different levels of linguistic abstraction. A mapping of elements in in the student and target answer that are not given in the question (in the following referred to as *new*) is established on different levels of abstraction on the token, chunk, and dependency triple (S. Bailey & Meurers, 2008). On the token level, the alignments are based on the linguistic evidence of lemma identity, spelling corrected identity, semantic type identity, synonym identity, and token identity. As described in (S. Bailey & Meurers, 2008), a set of alignment candidate configurations is given as input to the Traditional Marriage Algorithm (TMA, (Gale & Shapley, 1962)), which selects a global alignment configuration.

The third and final phase of the processing pipeline is the diagnosis step. In this step the alignments established in the previous step are quantified and given as input to a machine learning component that conducts a binary diagnosis of the semantic correctness of each student answer (Meurers, Ziai, Ott, & Kopp, 2011). For this purpose, the final component in the UIMA pipeline extracts a range of features that are given as input to the memory-based machine learner TiMBL (Daelemans, Zavrel, van der Sloot, & Van den Bosch, 2004). This module implements a k-nearest-neighbor algorithm with seven distance measures to conduct a majority voting of the diagnoses produced by the classifiers implementing different distance measures. Table 3.2 is adapted from (Meurers, Ziai, Ott, & Bailey, 2011, page 4) and (Meurers, Ziai, Ott, & Kopp, 2011, page 3) and lists the CoMiC baseline features. As shown in section 4 and section 5, the feature set will be augmented for the present work with general linguistic and task-specific alignment weighting features.

[1] http://opennlp.sourceforge.net/models-1.5/

Feature	Description
1. Keyword Overlap	Percent of keywords aligned
2. Target Token Overlap	Percent of aligned target tokens
3. Learner Token Overlap	Percent of aligned student tokens
4. Target Chunk Overlap	Percent of aligned target chunks
5. Learner Chunk Overlap	Percent of aligned student chunks
6. Target Triple Overlap	Percent of aligned target dependency triples
7. Learner Triple Overlap	Percent of aligned student dependency triples
8. Token Match	Percent of token alignments that were token-identical
9. Similarity Match	Percent of token alignments that were similarity-resolved
10. Type Match	Percent of token alignments that were type-resolved
11. Lemma Match	Percent of token alignments that were lemma-resolved
12. Synonym Match	Percent of token alignments that were synonym-resolved
13. Variety of Match (0-5)	Number of kinds of token-level alignments (features 8-12)

Table 3.2: CoMiC baseline features.

4 General Linguistic Weighting of Alignments

The first weighting discussed in this work measures quantities of general linguistic properties of the aligned elements in a student and in a target answer. This general linguistic weighting stands in contrast to the task-specific weighting proposed in section 5 with respect to the generality of the property encoded in the respective features. The motivation for the implementation of such features is as follows: given a student and a target answer with alignments between them that encode the similarity on different linguistic levels, to what degree are the aligned or new elements in the student and the target answer similar with respect to a general linguistic property not used before in the process of aligning elements? By adding a new dimension of similarity and quantifying this similarity on the elements in the student and target answer recognized to be similar on other linguistic levels of abstraction, the existing alignments are weighted and thereby assigned a relative importance with respect to a general linguistic property.

This approach distinguishes the resulting features from the CoMiC baseline features in that these features do not actively contribute to the alignment process, but rather take the alignment as given and assign a weight to these alignments. This approach has the advantage that it re-evaluates previously established alignments in a new dimension of similarity.

By adding features both for the student answer and the target answer, not only an unilateral quantity is measured, but the relation of the aligned elements of the two answers with respect to a general linguistic property is encoded. This allows the machine learning component to put the alignment quantities with respect to the general linguistic property into a relation and allows to detect overlaps and unmatched elements. For example suppose that in the student answer four new elements exhibit a certain linguistic property, but only one is aligned. In the target answer however only one element with this property is new and aligned. Features measuring the alignment quantity inside this group for the student and the target answer would output the value 1.0 for the target answer and 0.25 for the student answer, resulting in a skewed distribution of the general linguistic property of aligned elements in the two answers. This comparison is only possible via the means of abstraction from the surface form. Features measuring general linguistic properties thus should not rely on the surface form, but rather compare student and target answers with respect to quantities of equivalence classes that put the outcomes of the experiment of measuring the general

linguistic property on aligned or new elements into classes with similar behavior generally observed in the language under discussion.
In order to operationalize these considerations and test their effectiveness in practice, this work exemplifies the abstract concept of a general linguistic property by using part of speech classes as a means of determining a global kind of similarity of words in a language, and by using *tf.idf* weighting as a general concept with task-specific training data.
The following section will motivate the usage of the new features, and the subsequent subsections will explain the operationalization and implementation in more detail. The effect of these features on a short answer assessment system is tested and reported in the section 6.1.

4.1 Part of Speech Weighting of Alignments

Part of speech tags are labels assigned to tokens and represent syntactic, morphological, or semantic information about the corresponding token. They are an instance of a general linguistic property since the part of speech tag of a word can generally be assigned to a word in a language independent of an external context, as for example a concrete task (see section 5). Furthermore, parts of speech are a general linguistic concept applicable to every natural language (Mitkov, 2005).
What is however language-specific is the encoding of the linguistic properties in the part of speech tags. This is done via a part of speech tag set. The part of speech set used in this work is discussed in the next section.

4.1.1 STTS tag set

Since this work constitutes an extension of the CoMiC-DE system (Meurers, Ziai, Ott, & Kopp, 2011), the STTS tag set (Schiller, Teufel, & Thielen, 1995) as used by the baseline system is taken as the basis for this work.
As described in (Schiller et al., 1995), the tag set reflects distributional, syntactic, semantic, and morphological linguistic evidence. The tag set contains a total of 54 different tags. A full description of the tag set and the properties is given in (Schiller et al., 1995). This section focuses on the application of this tag set for alignment weighting.
Table 4.1 shows four groups of part of speech tags that are used in this work. The groups are a language-specific adaption of the part of speech groups first described in (Rudzewitz & Ziai, 2015) as an extension of the CoMiC-EN system.

Group	**STTS tags**
nominal	NN, NE
verbal	VVFIN, VVIMP, VVINF, VVIZU, VVPP, VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP
adjv	ADJA, ADJD, ADV
rest	APPR, APPRART, APPO, APZR, ART, CARD, FM, ITJ, KOUI, KOUS, KON, KOKOM, PDS, PDAT, PIS, PIAT, PIDAT, PPER, PPOSS, PPOSAT, PRELS, PRELAT, PRF, PWS, PWAT, PWAV, PAV, PTKZU, PTKNEG, PTKVZ, PTKANT, PTKA, TRUNC

Table 4.1: Coarse STTS subsets used for the general linguistic weighting.

The motivation for partitioning the tag set into these equivalence classes is the reduc-

tion of data sparsity by the means of linguistic knowledge. As discussed above, part of speech tags encode global properties in a language. This knowledge about the global similarity with respect to certain general property is used to form these groups. Even if the data contains only few instances with a certain part of speech tag, the groups will compensate for this sparsity by combining the frequencies of multiple similar part of speech tags and by normalizing this quantity.
One point of criticism for this approach is however that the group definitions lack an empirical foundation from which they can be inferred. Especially the 'rest' group contains many heterogeneous tags. The system may benefit from a model that takes this diversity into account and works with more subgroups. For this purpose, section 4.1.8 proposes an approach in which part of speech classes are automatically induced from the given data.

4.1.2 Implementation

Due to the exploratory character of this work, several different variants of part of speech features were implemented and tested. The discriminating element is the normalization of the different quantities, resulting in sets of feature variants expressing a different perspective on aligned and new elements each. The feature variants represent systematic variations of the formula described in (Rudzewitz & Ziai, 2015, page 2-3) with respect to the denominator. All features are implemented both for the student answer and for the target answer, resulting in two numerical values for each feature variant. All features return a value $v \in [0, 1], v \in \mathbb{R}$. For the explanation of the different feature variants it is important to keep in mind that only new elements not given in the question can be aligned, following an approach by (S. Bailey & Meurers, 2008). The features are implemented as feature extractor components in the final module of the CoMiC-DE UIMA pipeline (Meurers, Ziai, Ott, & Kopp, 2011). The following subsections will motivate and describe the different feature variants. Section 6 evaluates the performance of the short answer assessment system with different feature variants.

4.1.3 Terminology and Symbols

For the description of the feature variants the following notation is used: The basis for the computation is always a certain answer $A_h \in A$ from the set of all answers. Every answer contains a set of specific words $W_{A_h} \in W$ from the set of all words. Every word $w_j \in W_{A_h}$ in an answer is assigned a part of speech tag $t_{w_j} \in T$ from the part of speech tag set T for $0 < j < |W_{A_h}|$. The expression T_i denotes a subset of the part of speech tag set, for example the set of all verbal tags. For the computation of quantities a binary indicator function $[x]$ is defined where x is a truth conditional statement. The return value of the function is 1 if x is true, else the return value is 0. Formula 1 depicts this function.

$$[\,] : x \to \{0, 1\} \in \mathbb{N},\ [x] = \begin{cases} 1, & \text{if } x \text{ is true} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

For the combination of multiple atomic truth conditional statements, the operator AND serves as a representation of the logical conjunction operator. A complex truth

conditional statement consisting of two or more atomic expressions combined with AND is true if and only if every atomic statement of the complex expressions is true (Partee, Ter Meulen, & Wall, 2012, page 102).

4.1.4 Local Part of Speech Features

Local part of speech features express how many new words with one of a group of certain part of speech tag are aligned, normalized by the total number of new words with this part of speech tag. Intuitively spoken this feature gives an answer to the following question: Are many of the new tokens with this part of speech tag aligned ? Formula 2 shows the formal definition of a local part of speech feature.

$$\begin{aligned} ol_{local}(A_h, T_i) &= \frac{\#\text{ aligned tokens with pos } \in T_i}{\#\text{ new tokens with pos} \in T_i} \\ &= \frac{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [w_j \text{ is aligned AND } t_{w_j} = t \text{ AND } w_j \text{ is new}]}{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [t_{w_j} = t \text{ AND } w_j \text{ is new}]} \end{aligned} \tag{2}$$

4.1.5 Semi-Global Part of Speech Features

Semi-global part of speech features express how many new words with one of a group of certain part of speech tag are aligned, normalized by the total number of new words that are aligned. This feature variant answers the following question: What is the proportion of aligned tokens with a part of speech tag from this group with respect to all aligned tokens? Intuitively spoken this feature encodes whether many of the aligned tokens are from a certain part of speech group.
If the part of speech subsets form a partition of the part of speech tag set, then the whole set of semi-global part of speech features express a distribution of part of speech groups over the aligned tokens.
Formula 3 gives a formal definition of a semi-global part of speech feature.

$$\begin{aligned} ol_{sglobal}(A_h, T_i) &= \frac{\#\text{ aligned tokens with pos } \in T_i}{\#\text{ aligned tokens}} \\ &= \frac{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [w_j \text{ is aligned AND } t_{w_j} = t \text{ AND } w_j \text{ is new}]}{\sum_{w_j \in W_{A_h}} [w_j \text{ is aligned AND } w_j \text{ is new}]} \end{aligned} \tag{3}$$

4.1.6 Global Part of Speech features

Global part of speech features express how many new words with a tag from a group of certain part of speech tags are aligned, normalized by the total number of new words. These features express the proportion of aligned tokens with one of the tags of this group with respect to all new tokens. Intuitively such features answer the following question: do many of the new words have a tag from this part of speech group and are at the same time aligned ?
Formula 4 gives a formal definition for a global part of speech feature.

$$
\begin{aligned}
ol_{global}(A_h, T_i) &= \frac{\#\text{ aligned tokens with pos } \in T_i}{\#\text{ new tokens}} \\
&= \frac{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [w_j \text{ is aligned AND } t_{w_j} = t \text{ AND } w_j \text{ is new}]}{\sum_{w_j \in W_{A_h}} [w_j \text{ is new}]}
\end{aligned} \tag{4}
$$

4.1.7 Interpolated Part of Speech features

Interpolation is defined in the context of this work as the combination of different feature variants. Two interpolation variants were implemented, resulting in a set of features each. This work focuses on linear interpolation and non-linear interpolation with equal weights each.

Non-Linear Interpolation Non-linear interpolation is obtained by multiplying the respective output values from the local, semi-global, and global feature variants. This feature variant punishes zero quantities for one feature variant with the means of the multiplication operation.
Formula 5 shows a formal definition of a non-linear interpolated feature.

$$
ol_{ip}(A_h, T_i) = ol_{local}(A_h, T_i) \times ol_{sglobal}(A_h, T_i) \times ol_{global}(A_h, T_i) \tag{5}
$$

Linear Interpolation Linear interpolation features represent a linear combination of the local, semi-global, and global feature variant values. Each of these three variants is assigned an equal weight of $\frac{1}{3}$. This feature variant is not as sensitive to zero quantities as the non-linear interpolation feature variant.
Formula 6 gives a formal definition of a linear interpolation feature.

$$
ol_{lip}(A_h, T_i) = \frac{1}{3} \times (ol_{local}(A_h, T_i) + ol_{sglobal}(A_h, T_i) + ol_{global}(A_h, T_i)) \tag{6}
$$

4.1.8 Alternative Part of Speech Group Determination

One problem concerning the implementation of part of speech features is the determination of equivalence classes of part of speech tags. This generalization constitutes a step towards on the one hand reducing data sparsity, and on the other hand avoiding overfitting of the data. Basically two approaches are possible for solving this problem. The approaches are (1) using linguistic knowledge to form the classes and (2) inferring the classes from the data without linguistic knowledge. While (1) constitutes a top-down approach, approach (2) is a bottom-up data-driven process to be tackled by unsupervised machine learning techniques. Section 4.1.1 presented the first approach, and this section presents the second route.

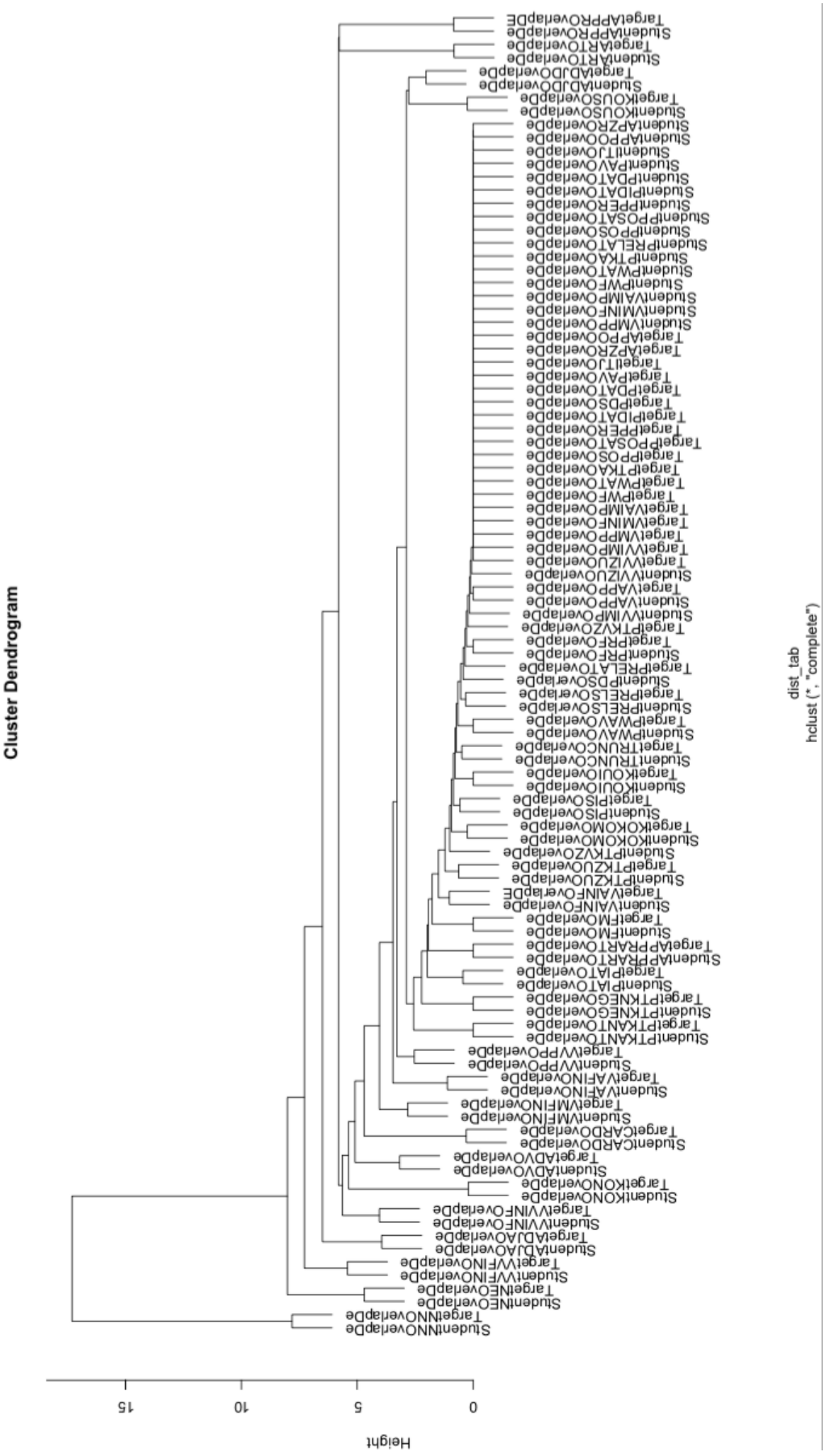

Figure 4.1: Hierarchical Agglomerative Clustering of Part of Speech Tags over all instances of CREG-1032.

Figure 4.1 shows a cluster of part of speech tag alignment quantities over all instances of CREG-1032. For the purpose of this illustration, the local feature computation variants (see section 4.1.4) were used. For every instance, every part of speech feature was computed. The resulting matrix was loaded into R and a hierarchical, bottom-up agglomerative cluster was created with the *hclust* function[3].
The resulting dendrogram depicted in figure 4.1 shows the similarity of part of speech tags over all instances. Longer branches indicate a bigger distance between the respective elements.
The dendrogram shows that for the majority of the part of speech features, the student and target variant of one feature are most similar. The result is a considerable amount of clusters representing one part of speech tag feature in its two variants. Another observation is that there is one big cluster of very similar tags. This cluster starts at the center of the dendrogram and includes all tags up to the *StudentKOUSOverlapDe* feature. The data shows that this cluster resulted from zero values of the respective part of speech features.
Apart from that, several part of speech features are more distinct from the rest. Interestingly the part of speech features corresponding to these exposed clusters are exactly representatives of the main word classes described in (Schiller et al., 1995). In detail the dendrogram shows that there is a difference in global behavior with respect to alignment quantities for the word classes nouns (*NN*, *NE*), verbs (*VVFIN*, *VVINF*, *VMFIN*, *VAFIN*, *VVPP*), adjectives (*ADJA*, *ADJD*), adverbs (*ADV*), numbers (*CARD*), conjunctions (*KON*), articles (*ART*), particles(*PTKA*, *PTKNEG*), and appositions (*APPRART*). This patterns is remarkable since the clustering as an unsupervised machine learning algorithm is provided with no assumptions at all about linguistic word classes. This approach thus shows that there is in fact a difference in alignment preferences with respect to the general linguistic property of part of speech classes that can be derived from the data. In order to test whether features reflecting this insight improve the system, an evaluation is conducted and reported in section 4. Since the part of speech features for the distinct clusters can be seen as representatives of the STTS main word classes (Schiller et al., 1995), and since the part of speech features are implemented with an intrinsic normalization, the effectiveness of features reflecting these word classes is tested by implementing the STTS main word classes as part of speech features. Thereby the discussed problem of data sparsity and the problem of deciding where exactly to cut the cluster into sub-clusters for new features are overcome. Table 4.2 shows the part of speech groups representing the main word classes in the STTS tag set (Schiller et al., 1995) as used in the present work. The punctuation class is left out since punctuation is filtered out by the system and since memory-based approaches are sensitive to irrelevant features (Hall et al., 2009).

4.2 tf.idf Weighting of Alignments

An alternative instance of a general linguistic weighting measure is the *term frequency inverse document frequency* (in the following abbreviated as *tf.idf*) weighting scheme. The *tf.idf* is a measure originating from the area of information retrieval.
This approach implements a general linguistic measure inasmuch as the *tf.idf* weighting is not specific to a reading comprehension task, but can be used for the weighting of arbitrary terms in an arbitrary collection of documents in any language. The realiza-

[3] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html

Group	STTS tags
nouns	NN, NE
adjectives	ADJA, ADJD
numbers	CARD
verbs	VVFIN, VVIMP, VVINF, VVIZU, VVPP, VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP
articles	ART
pronouns	PPER, PRF, PPOSAT, PPOS, PDAT, PDS, PIDAT, PIS, PIAT, PRELAT, PRELS, PWAT, PWS, PWAV, PAV
adverbs	ADV
conjunctions	KOUI, KOUS, KON, KOKOM
appositions	APPR, APPRART, APPO, APZR
particles	PTKZU, PTKNEG, PTKVZ, PTKA, PTKANT
other	ITJ, TRUNC, XY, FM

Table 4.2: STTS main word classes used for the general linguistic weighting.

tion of this measure however takes into account the specific task context of a student answer since the training data for this measure consists of the reading comprehension text based on which the reading comprehension tests are conducted. This approach can thus be seen as a hybrid approach between general linguistic and task-specific alignment weighting.

4.2.1 Additional Symbols and Terminology

In addition to the symbols defined in section 4.1.3, the following symbols are used in the definition of the *tf.idf* weighting features. $d_i \in D$ is a document in the collection of reading comprehension texts (documents) D represented as a bag of lemmas. The usage of lemmas instead of surface forms is motivated by the need for the reduction of data sparsity. Given a token w_j, d_i is the document that the question being answered was posed to. This means that given a student answer, the corresponding document used in the *tf.idf* measure is the reading comprehension text used a basis for the student answer.
OR denotes the logical inclusive or operator. An expression consisting of two atomic expressions combined with OR is true if one of the expressions is true or if both of them are true (Partee et al., 2012, page 103). $tf_{j,i}$ denotes the number of occurrences of the lemma of the token w_j in the bag of lemmas of the document d_i. This value indicates the term frequency of a lemma of word w_j from the student answer in the corresponding reading comprehension text. df_j is the document frequency of the lemma of the word w_j. It indicates how many documents (reading comprehension texts) exist such that the set of lemmas of words in this document contain the lemma of w_j. N simply indicates the total number of documents in the collection of reading comprehension texts.

4.2.2 Formal Feature Definition

The quantities described in the last section are combined in the *tf.idf* formula shown in formula 7 and 8. This formula is based on the formula from (Manning & Schütze, 1999, page 543). It was adapted to process information about the alignment and givenness status of input words.

$$ol_{tf.idf}(A_h) = \sum_{w_j \in W_{A_h}} weight_{tf.idf}(w_j, d_i) \tag{7}$$

with

$$weight_{tf.idf}(w_j, d_i) = \begin{cases} 0 & \text{, if } (w_j \text{ NOT new}) \text{ OR} \\ & (w_j \text{ NOT aligned}) \text{ OR} \\ & (w_j \notin d_i) \\ (1 + log(tf_{j,i})) \times log\frac{N}{df_j} & \text{, otherwise} \end{cases} \tag{8}$$

For every aligned token, the *tf.idf* value is measured and summed for all tokens of this sentence. This measurement is conducted for the student and target answer, resulting in two numeric features. The features indicate the informativeness of the aligned words in this specific task context.

4.2.3 Feature Motivation

tf.idf measures have been widely described in the information retrieval literature (e.g. (Salton & Buckley, 1988), (Mitkov, 2005), (Manning & Schütze, 1999)). Therefore this section focuses on the motivation for this measurement in the specific domain of short answer assessment.

The *tf.idf* formula as defined in the previous section is a product consisting of two parts: the first part represents the smoothed term frequency. It transforms the frequency of a term in the current document onto a logarithmic scale and adds this value to 1. With this part of the formula, the frequency of a term in a document is taken into account. The second part of the product puts the term frequency into a relation in the context of other documents. It basically measures how many of the other documents also contain this term. If the term occurs in also approximately every other document in the collection, then it is presumably less informative for describing the content of this document.

In this context the *tf.idf* measure is used to wight the importance of terms in the student answer by using the task context in the form of the reading comprehension text based on which this answer was given. All reading comprehension texts from the CREG-5K-OSU and CREG-5K-KU are taken, and *tf.idf* values are computed for the lemma of every token in this collection, resulting in two indexes from lemmas to documents to *tf.idf* values. These values are used to determine the weight of the lemmas of the aligned elements in the student answer. If a word is important in the corresponding reading text, then the alignment of this word in the student answers contributes important information and a high *tf.idf* value is added to the feature return value.

5 Task-Specific Weighting of Alignments

The task-specific weighting of alignments represents a more specialized approach of alignment weighting in that it takes into account the specific task context in which a student answer was produced.

Previous publications about the CREG corpus (e.g. (Meurers et al., 2010), (Ott et

al., 2012)) emphasized on the importance of taking into account the task associated with a short answer. Therefore, the CREG corpus contains not only student responses to short answer questions, but also the questions themselves, target answers to every question, the corresponding reading text, and meta data about the different elements. While the definition of the term *task* includes many aspects from various research areas related to educational research and is discussed controversially (Ellis, 2009), the present work only focuses on a certain aspect of the characterization of a task. This work exemplifies the inclusion of the task context into a short answer assessment system by encoding the question type of a question to which a student answer was given, thereby continuing previous research (Meurers, Ziai, Ott, & Kopp, 2011) on this aspect.

The implementation of a task-based alignment weighting feature is realized by encoding subtypes (see (Meurers, Ziai, Ott, & Kopp, 2011)) of the short answer question type (see (Burrows et al., 2015)). The following paragraphs will first define the abstract question type of a short answer question before the more specific subtypes will be discussed in detail.

The abstract type of question encoded in CREG is the *short answer question* type (Burrows et al., 2015, page 1-2). They name five criteria that define a short answer question that will be listed in the following.

The first criterion is that the answer to a question is not inferable from the question itself, but external knowledge is needed to answer a question. This aspect is fulfilled since the corpus only contains questions whose answer is explicitly encoded in the text (Meurers et al., 2010).

The second criterion requires the answer to be in natural language. This is also the case for the CREG corpus, since both target and student answers are represented as written natural language. The language in the student answers can be seen as an inter-language deviating from the standard language as produced by native speakers and representing an intermediate state of proficiency. Thus the language is a natural learner language.

This leads to the third criterion proposed by (Burrows et al., 2015), namely the restriction of the length of the answers to between one phrase and one paragraph. (Ott et al., 2012) analyzed the average length of student responses on the representative sample of CREG-1032. They measured an average length of 5.4 tokens per answer (KU data set) and 11.4 tokens per answer (OSU data set). The length requirement is thus also fulfilled.

The fourth criterion pertains the grading of the responses. (Burrows et al., 2015) put emphasis on the fact that the grading be conducted on the basis of the meaning of the student answer, and not on the basis of the form of the answer. (Ott et al., 2012) explicitly state that form errors are not regarded in the process of conducting the binary diagnosis. This results in a "significant variation in form, including a high number of form errors" (Meurers, Ziai, Ott, & Kopp, 2011, page 1).

The fifth criterion requires the restriction of the "level of openness" (Burrows et al., 2015, page 2) of the answer by the question. This means that the space of possible inputs is restricted with questions that require responses whose content is clearly defined. As stated above, all questions ask about knowledge explicitly given in the text. The expression of these concepts, however, opens up the space of possible inputs and thus the openness to a certain degree. This issue will be discussed in the next paragraph.

(Meurers, Ziai, Ott, & Kopp, 2011) use the question type taxonomy for comprehension questions proposed by (Day & Park, 2005) to categorize the question types in the CREG-1032 corpus and measure the system performance on different question types. They distinguish the following ten question forms: *Alternative*, *How*, *What*, *When*, *Where*, *Which*, *Who*, *Why*, *Yes/No*, *Several*. These categories are not purely surface-based, since multiple surface question forms can be mapped to the same question form. For example the question words *Woraus* and *Womit* differ in their surface form, but are both mapped to the question type *What*. (Meurers, Ziai, Ott, & Kopp, 2011) showed that the CoMiC-DE system's performance varied across question types. Question forms that allow a wider input space in terms of the surface expression of the semantic concepts in the answer were found to pose higher difficulties to the system, resulting in a lower accuracy in the binary classification task. They found out that especially the questions with a *why* or a *who* question form could not be classified as accurately by the CoMiC-DE system as the other question types. Another problematic case reported by this article is the *alternative* question form. Since the CoMiC system excludes material given in the question from the alignment process, the alternatives given in the question can't be aligned with material from the student answer.
This variation in system performance given a certain question type motivated the implementation of an alignment weighting strategy that takes into account the question form of the question to which the student answer currently to be diagnosed was given to. Section 6.2 reports on the performance of CoMiC-DE when extended with features encoding the weighting of alignments by question type, and section 6.3 shows results for combining this weighting method with the general linguistic weighting proposed in section 4.

6 Experimental Testing of the Approach

This section reports on the effect of the alignment weighting methods presented in section 4 and 5 on the performance of the CoMiC-DE system. The features are implemented and the performance of the CoMiC-DE system is measured. This extrinsic, experimental testing approach ensures that the effects are relevant in practice. Factors that can have an influence on the experimental results are reported in section 6.1.1. Results are reported separately for all different combinations of factors, such as the data set or the feature normalization variant.
The structure of this section is as follows: Section 6.1 tests the effect of a general linguistic weighting of alignments before section 6.2 reports on the effects of a task-specific weighting of alignments. Section 6.3 finally reports on the effect of combining the two weighting methods. In addition, section 6.3.4 and section 6.3.5 present analyses of the effect of the new features with respect to the different question types.

6.1 General Linguistic Weighting of Alignments

6.1.1 Introduction

This section presents experimental results for the extension of an alignment-based short answer assessment system with part of speech alignment and *tf.idf* weighting features. The following hypotheses are tested: the binary classification performance

of an alignment-based short answer assessment system increases if this system is augmented with features that weight alignments dependent on the part of speech of the aligned elements or features that weight the alignments with *tf.idf* features.
The results depend on various factors that need to be taken into consideration in order to objectively evaluate the impact of the new features. The following paragraphs will discuss the variables that need to be controlled in the experiment.
As proposed in section 4, different feature variants are explored. The feature variants have an influence on the setup and are thus controlled for by providing results for all individual feature variants. Furthermore, the groups of part of speech features that are implemented via the different feature variants play a role in the results. Test are conducted for setups with four coarse groups (nominal ,verbal, adjective/adverbial, rest), as well as with the main word classes from the STTS tag set.
Another important factor for the experiment is the data that is used for the evaluation. In order to test whether the part of speech features generalize and can be successfully applied for different data sets, the results section presents results for various data sets of CREG with different characteristics. Furthermore, the system is evaluated on all subset sets of CREG for which results have been published previously ((Meurers, Ziai, Ott, & Kopp, 2011), (Horbach et al., 2013), (Burrows et al., 2015), (Hahn & Meurers, 2012)) in order to make the results comparable to related work.
Another factor influencing the interpretation of the results is the baseline system used for the experiment. As discussed in section 3, some tools and models were updated. Also the selection of a target answer when none is annotated changed from taking the first one to taking the one with the smallest Levenshtein distance (Levenshtein, 1966) to the corresponding student answer. This led to an improvement in the baseline system performance on the CREG-1032-OSU data set from 84.6% (Meurers, Ziai, Ott, & Kopp, 2011) to 87.0% in the present work.
A further crucial point in the evaluation is the choice of a machine learning algorithm and its implementation. To make the work comparable to previous studies (Meurers, Ziai, Ott, & Kopp, 2011), (Meurers, Ziai, Ott, & Bailey, 2011), (S. M. Bailey, 2008), results are reported with the TiMBL memory-based learner (Daelemans et al., 2004). The same setup as for previous publications (S. M. Bailey, 2008),(Meurers, Ziai, Ott, & Kopp, 2011) is used, namely a k=1-nearest neighbor leave-one-out testing with a majority voting over seven classifiers, each employing a different distance measure (dot product, cosine, overlap, numeric overlap, Levenshtein, modified value difference, Jeffrey divergence).
The next section shows the formal approach for testing these hypotheses.

6.1.2 Methods

The formal representation of the hypotheses and factors above that are to be tested is as follows:

H_0: The binary classification performance of an alignment-based short answer assessment system does not change if it is augmented with part of speech or *tf.idf* features.

H_1: The binary classification performance of an alignment-based short answer assessment system significantly improves if it is augmented with part of speech or *tf.idf* features.

These hypotheses are tested for every data set and for every feature set. In order to test whether there is a significant improvement that is not due to chance, a statistical significance test is conducted for every of the above named conditions. For every data

set and test, the same baseline system is used.
The test used for objectively testing the results is the McNemar test. This statistical significance test is used due to the following reasons (see (Gries, 2008), (Dietterich, 1998)). The data is not normally distributed, but follows a binomial distribution, since for every answer the prediction can be either true or false. Furthermore, the samples are not independent. Both the baseline system and the augmented system use the same evaluation data. In addition, the augmented system represents an extension of the baseline system, making the systems' predictions themselves dependent. The data consists of paired categorical outcomes (baseline versus augmented system prediction), making it a natural input to the McNemar test.
The significance testing is conducted in R[4](Ihaka & Gentleman, 1996) . The significance level was set to $\alpha = 0.05$ before the experiments were conducted. Since the McNemar test is based on the χ^2 distribution, the test outcomes are reported in this metric. The null hypothesis is rejected if and only if the test outcome is bigger than the critical value 3.84 for $\alpha = 0.05$ and 1 degree of freedom.
The next section presents the quantitative results of the experiments.

6.1.3 Results

Coarse Part of Speech features. Table 6.1 shows the quantitative results of the experiments for the CoMiC-DE system augmented with parts of speech features implementing the four coarse part of speech classes defined in section 4.1.1.

system variant	3620-KU	3620-OSU	1032-KU	1032-OSU	5K-KU	5K-OSU
baseline	81.5	82.2	84.6	87.0	80.9	82.5
local	82.0	82.6	85.2	**90.0***	82.0	82.8
semi-global	81.2	**84.1***	85.4	87.2	81.3	**84.0***
global	83.0	**83.6***	84.8	85.8	81.6	**83.6***
ip	80.5	**84.1***	85.1	85.1	81.7	**84.4***
lip	82.6	**84.1***	84.4	87.0	81.4	**84.1***

Table 6.1: System performance for the baseline system augmented with part of speech features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

In nine cases there is a statistically significant improvement. In these cases the null hypothesis H_0 is rejected. Table 6.1 shows that every feature variant can improve the system performance significantly. This observation is made across different data sets. It is worth noticing that all the significant improvements here were obtained on the OSU data sets.
The strongest statistically significant improvement over the baseline was reached on the CREG-1032-OSU data set. For this data set the baseline is already by far the strongest with 87.0% accuracy, but the local part of speech features increase the performance by another 3.0%, resulting in an accuracy of 90.0%. This is the highest value obtained in any of the conducted experiments.
For the CREG-3620-OSU and the CREG-5K-OSU data set, the widest range of feature variants resulted in a statistically significant improvement. The semi-global, global, interpolated, and linear interpolated feature variants significantly improved the system

[4]https://www.r-project.org/

performance.

Table 6.2 provides the test statistics used in the hypothesis testing for the statistically significant results. The table shows the obtained χ^2 test statistics as well as the corresponding p-values. This table is provided in order to enable later comparisons of system performance to this work.

Data set	**feature variant**	χ^2	**p-value**	**decision**
CREG-1032-OSU	local	6.2593	0.01235	reject H_0
CREG-5K-OSU	semi-global	7.8125	0.005189	reject H_0
CREG-5K-OSU	global	3.9755	0.04617	reject H_0
CREG-5K-OSU	ip	11.81272	0.0005883	reject H_0
CREG-5K-OSU	lip	7.815	0.005181	reject H_0
CREG-3620-OSU	semi-global	10.7651	0.001034	reject H_0
CREG-3620-OSU	global	5.88	0.01531	reject H_0
CREG-3620-OSU	ip	11.2	0.000818	reject H_0
CREG-3620-OSU	lip	9.6556	0.001888	reject H_0

Table 6.2: Hypothesis test statistics for cases of statistically significant system performance improvement for the baseline system augmented with part of speech features.

STTS Part of Speech Features. Section 4.1.8 motivated the use of the coarse part of speech classes provided by the STTS tag set. Table 6.3 shows the results for a system augmented with part of speech features based on STTS main word groups (see (Schiller et al., 1995)). Table 6.4 provides the corresponding test statistics for the cases in which the null hypothesis was rejected.

system variant	**3620-KU**	**3620-OSU**	**1032-KU**	**1032-OSU**	**5K-KU**	**5K-OSU**
baseline	81.5	82.2	84.6	87.0	80.9	82.5
stts-local	83.1	83.1	86.1	88.6	82.1	83.1
stts-semi-global	83.0	**84.2***	85.6	87.2	81.8	**84.5***
stts-global	82.6	**83.6***	86.4	87.9	82.0	**83.7***
stts-ip	81.5	83.0	84.3	86.7	81.8	**84.3***
stts-lip	82.2	**84.4***	86.2	86.7	81.8	**84.9***

Table 6.3: System performance for the baseline system augmented with STTS group part of speech features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Data set	**feature variant**	χ^2	**p-value**	**decision**
CREG-3620-OSU	stts-semi-global	11.0887	0.0008685	reject H_0
CREG-3620-OSU	stts-global	5.6903	0.01706	reject H_0
CREG-3620-OSU	stts-lip	13.162	0.0002857	reject H_0
CREG-5K-OSU	stts-semi-global	12.7644	0.0003533	reject H_0
CREG-5K-OSU	stts-global	4.768	0.02899	reject H_0
CREG-5K-OSU	stts-ip	10.7771	0.001028	reject H_0
CREG-5K-OSU	stts-lip	18.0899	$2.107e-05$	reject H_0

Table 6.4: Hypothesis test statistics for the baseline system augmented with STTS group part of speech features for cases of statistically significant system performance improvement.

In seven cases a statistically significant improvement of the system performance was measured. As in the experiments with the coarse part of speech groups, statistically

significant improvements could only be obtained on the OSU data sets. The STTS group features could improve only for configurations of data sets and feature variants where the coarse part of speech groups also yielded a statistically significant improvement. In cases where the STTS group features yielded a statistically significant improvement, the improvements are marginally higher as with the coarse part of speech groups. However, the STTS group features did not improve the performance significantly in two cases where the coarse part of speech features made this improvement.

tf.idf features. Table 6.5 shows the effect of the second general linguistic weighting method, the *tf.idf* weighting. The two *tf.idf* features improve the accuracy of the system on every data set. Four out of six improvements are statistically significant (see table 6.6).

system variant	3620-KU	3620-OSU	1032-KU	1032-OSU	5K-KU	5K-OSU
baseline	81.5	82.2	84.6	87.0	80.9	82.5
tf.idf	**84.2***	**84.1***	86.1	88.4	**83.1***	**84.3***

Table 6.5: System performance for the baseline system augmented with *tf.idf* features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Data set	feature variant	χ^2	p-value	decision
CREG-5K-KU	tf.idf	9.3023	0.002289	reject H_0
CREG-5K-OSU	tf.idf	11.7622	0.0006044	reject H_0
CREG-3620-OSU	tf.idf	11.1298	0.0008495	reject H_0
CREG-3620-KU	tf.idf	5.7143	0.01683	reject H_0

Table 6.6: Hypothesis test statistics for the baseline system augmented with *tf.idf* features for cases of statistically significant system performance improvement.

6.2 Task-Specific Weighting of Alignments

6.2.1 Introduction

In this section the effectiveness of a task-specific weighting of alignments is tested. As an instance of a task-specific weighting, the question types of questions to which student answers were given are implemented. As discussed in the previous section, different data sets have different characteristics and pose different problems to the system. This represents thus a factor influencing the results. For this reason, results are reported separately for every data set.

6.2.2 Methods

The following hypotheses are tested:

H_0: The binary classification performance of an alignment-based short answer assessment system does not change if it is augmented with question type features.

H_1: The binary classification performance of an alignment-based short answer assessment system significantly improves if it is augmented with question type features.

In order to test the hypotheses eleven new features are added to the CoMiC baseline features. Each feature represents the output of a binary indicator function indicating the presence of a certain question form in the question the current student answer was given to. Eleven features are implemented for ten question forms since one feature is used as a fallback if no question form could be determined.
In order to evaluate the validity of the basis of the features, the question type detection approach was tested on CREG-1032, since this data set contains gold annotations for the discussed question forms from a previous study (Meurers, Ziai, Ott, & Kopp, 2011). Section 6.2.3 discusses this component of the system.

6.2.3 Results

Table 6.7 shows the results of the experiments with the CoMiC-DE system in terms of accuracy when augmented with question type features. Only one improvement is statistically significant. In one case, the performance of the system decreased. As shown in section 6.3.4, one possible explanation for this is the very uneven distribution of question types in the data and the equally imbalanced distribution of binary diagnosis labels given a certain question type.

system variant	**3620-KU**	**3620-OSU**	**1032-KU**	**1032-OSU**	**5K-KU**	**5K-OSU**
baseline	81.5	82.2	84.6	87.0	80.9	82.5
q-types	80.8	**83.1***	85.4	87.2	80.9	82.8

Table 6.7: System performance for the baseline system augmented with question type features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Data set	**feature variant**	χ^2	**p-value**	**decision**
CREG-3620-OSU	q-types	6.068	0.01377	reject H_0

Table 6.8: Hypothesis test statistics for the baseline system augmented with question type features for cases of statistically significant system performance improvement.

Question Type Detection. An automatic question type detection module was implemented for this work. From a previous study (Meurers, Ziai, Ott, & Kopp, 2011), gold standard annotations were available for the CREG-1032-OSU and the CREG-1032-KU data set. These data sets were used to optimize the question type detection module.
For 165 out of 175 different questions associated with student answers, the question type detection module predicted the correct question form. The module thus predicted the correct question form in 94% of all cases. This number has however to be treated with caution, since due to the limited amount of data, no out-of-domain testing of the performance of the question type detection module could be conducted.

Error Analysis. A manual inspection of the ten cases where the question type detector's predictions deviated from the gold standard annotation showed various sources of errors. The most frequent mismatch occurred when the system predicted the question form *what* when the gold label was *which*. The human annotators presumably used more of the task context of this question. If the reading comprehension text associated with this question contains a set of alternatives of which the question asks for one, the gold label was set to *which* for this question. In contrast to that, the question type module did not incorporate the reading comprehension text in the process of determining the question form. Thus it did not refer to the limited set of alternatives and assumed that the question asks for any object, making the *what* question type a natural choice.
Another mismatch concerns the distinction between one atomic question type and the *several* question type. The gold annotation partly specified an atomic question type for questions where two surface form question occurred. If the question asks for two aspects of a certain object, for example the advantages and disadvantages of this object, the gold annotation specified the question type *what*. The question type detection module in contrast predicted the question form *several* since it treated the question asking for the advantages and the question asking for the disadvantages as two different questions.

6.3 Combination of General Linguistic and Task-Specific Alignment Weighting

6.3.1 Introduction

This section presents results in terms of the accuracy of the CoMiC-DE system when augmented with a combination of general linguistic and task-specific alignment weighting methods. The motivation for this combination is to test whether there are question type specific part of speech alignment patterns that can be used by the system to improve the classification accuracy. Another hypothesis to be tested is whether the *tf.idf* weighting trained on the reading comprehension texts can be effectively combined with the question type features. Although *tf.idf* represents a general linguistic property, in this context it is used to encode the task context from the perspective of the text, which is combined with the question types, which represent the task context from the perspective of the question. Finally this section presents experiments for combining all three weighting methods. In addition to that, an analysis of the most informative part of speech alignments per question type is provided, together with an analysis of the question type specific system performance (macro-averages).
As motivated in the previous sections, the results are reported for different data sets.

6.3.2 Methods

The following hypotheses are tested:

H_0: The binary classification performance of an alignment-based short answer assessment system does not change if it is augmented with question type and/or *tf.idf* and/or part of speech features.

H_1: The binary classification performance of an alignment-based short answer assessment system significantly improves if it is augmented with question type and/or *tf.idf* and/or part of speech features.

6.3.3 Results

Question Types and STTS Part of Speech Features Table 6.9 shows the accuracy of the short answer assessment system when augmented with the question type and part of speech weighting (STTS groups) features. In table 6.10, the corresponding test statistics are shown for cases of statistically significant improvements. In seven out of 30 cases, there is a statistically significant improvement. More improvements can be observed, but they are not significant from a statistical point of view. In comparison to other combinations of weighting methods, as for example shown in table 6.17, the number of statistically significant improvements is rather low. In order to understand this observation, an analysis of the importance of part of speech features was conducted to see which part of speech features are actually most important given a certain question type. Section 6.3.4 presents this analysis.

system variant	**3620-KU**	**3620-OSU**	**1032-KU**	**1032-OSU**	**5K-KU**	**5K-OSU**
baseline	81.5	82.2	84.6	87.0	80.9	82.5
q-types + stts local	82.0	**83.6***	87.2	87.7	81.9	83.6
q-types + stts semi-global	82.6	**84.1***	85.2	86.3	81.2	**84.4***
q-types + stts global	82.0	**83.7***	86.9	86.5	81.1	83.6
q-types + stts ip	81.0	**83.7***	86.7	86.0	81.1	84.0
q-types + stts lip	81.5	**83.8***	87.0	85.5	81.5	**84.4***

Table 6.9: System performance for the baseline system augmented with question type and STTS group part of speech features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Data set	**feature variant**	χ^2	**p-value**	**decision**
CREG-3620-OSU	q-types + stts local	4.7337	0.02958	reject H_0
CREG-5K-OSU	q-types + stts semi-global	10.5279	0.001176	reject H_0
CREG-3620-OSU	q-types + stts semi-global	9.0299	0.002656	reject H_0
CREG-3620-OSU	q-types + stts global	5.7601	0.01639	reject H_0
CREG-3620-OSU	q-types + stts ip	5.9838	0.01444	reject H_0
CREG-3620-OSU	q-types + stts lip	6.9818	0.008234	reject H_0
CREG-5K-OSU	q-types + stts lip	10.2507	0.001366	reject H_0

Table 6.10: Hypothesis test statistics for the baseline system augmented with question type and STTS group part of speech features for cases of statistically significant system performance improvement.

Question Types plus tf.idf Weighting In table 6.11 the results for a system that combines question types and *tf.idf* measures to weight the alignments is shown. In table 6.12 the test statistics are shown for cases of statistically significant improvements. The combination of weighting alignments by question type and weighting alignments by *tf.idf* measures yields an improvement in every case, of which four are statistically significant. For the CREG-3620 data sets, the combination of both weighting methods could improve the performance when compared to the single weighting methods.

system variant	3620-KU	3620-OSU	1032-KU	1032-OSU	5K-KU	5K-OSU
baseline	81.5	82.2	84.6	87.0	80.9	82.5
q-types + tf.idf	**84.4***	**84.5***	86.1	87.0	**82.4***	**84.5***

Table 6.11: System performance for the baseline system augmented with question type and *tf.idf* weighting in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Data set	feature variant	χ^2	p-value	decision
CREG-5K-OSU	q-types + tf.idf	14.718	0.0001248	reject H_0
CREG-5K-KU	q-types + tf.idf	4.3077	0.03794	reject H_0
CREG-3620-KU	q-types + tf.idf	6.3913	0.01147	reject H_0
CREG-3620-OSU	q-types + tf.idf	16.1678	$5.797e-05$	reject H_0

Table 6.12: Hypothesis test statistics for the baseline system augmented with question type and *tf.idf* weighting for cases of statistically significant system performance improvement.

STTS Part of Speech Features plus tf.idf Table 6.13 shows how the system performs when augmented with part of speech (STTS groups) features and *tf.idf* features for the weighting of alignments. In table 6.14 the corresponding test statistics are shown. The combination of part of speech (STTS group) and *tf.idf* features yields better results than the combination of part of speech features and question type features. For this feature set there are 17 cases of statistically significant improvements. These improvements are however less equally distributed across the data sets as for the combination of all feature variants (see table 6.17). This feature variant works very well for the CREG-5K-OSU and CREG-3620-OSU data set. The improvements are not restricted to the OSU data sets, since there are also significant improvements for the 5K-KU, the 3620-KU, and the 1032-KU data sets.

Coarse Part of Speech Features plus tf.idf Table 6.15 shows the system performance for the system when the four coarse part of speech classes are used (in contrast to the STTS groups) and the alignments are additionally weighted with *tf.idf* weights. Table 6.16 shows the test statistics for the 15 cases in which a statistically significant improvement was measured. The combination of coarse part of speech feature classes and *tf.idf* weighting works best for CREG-5K-KU and the CREG-5K-OSU data set. For CREG-3620-OSU there are also significant improvements, parallel to other feature sets.

system variant	3620-KU	3620-OSU	1032-KU	1032-OSU	5K-KU	5K-OSU
baseline	81.5	82.2	84.6	87.0	80.9	82.5
stts local + tf.idf	**83.8***	**85.3***	87.0	87.0	82.3	**85.0***
stts semi-global+ tf.idf	**83.9***	**84.7***	86.2	87.7	72.4	**84.8***
stts global+ tf.idf	83.5	**84.4***	**88.0***	85.3	**82.8***	**84.6***
stts ip+ tf.idf	83.3	**84.5***	86.1	86.3	**82.6***	**84.9***
stts lip+ tf.idf	**84.5***	**84.5***	86.7	85.8	**83.0***	**84.7***

Table 6.13: System performance for the baseline system augmented with STTS group part of speech and *tf.idf* weighting in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Data set	feature variant	χ^2	p-value	decision
CREG-3620-KU	stts lip+ tf.idf	3.967	0.0464	reject H_0
CREG-3620-KU	stts local+ tf.idf	5.2609	0.02181	reject H_0
CREG-3620-KU	stts semi-global+ tf.idf	3.6	0.05778	reject H_0
CREG-3620-OSU	stts global+ tf.idf	11.4499	0.000715	reject H_0
CREG-3620-OSU	stts ip+ tf.idf	12.1676	0.0004863	reject H_0
CREG-3620-OSU	stts lip+ tf.idf	13.8443	0.0001986	reject H_0
CREG-3620-OSU	stts local+ tf.idf	22.7528	$1.842e-06$	reject H_0
CREG-3620-OSU	stts semi-global+ tf.idf	14.3204	0.0001542	reject H_0
CREG-1032-KU	stts global+ tf.idf	6.4516	0.01109	reject H_0
CREG-5K-KU	stts global+ tf.idf	5.303	0.02129	reject H_0
CREG-5K-KU	stts ip+ tf.idf	4.4286	0.03534	reject H_0
CREG-5K-KU	stts lip+ tf.idf	6.3333	0.01185	reject H_0
CREG-5K-OSU	stts global+ tf.idf	11.3333	0.0007613	reject H_0
CREG-5K-OSU	stts ip+ tf.idf	16.8066	$4.139e-05$	reject H_0
CREG-5K-OSU	stts lip+ tf.idf	13.5147	0.0002367	reject H_0
CREG-5K-OSU	stts local+ tf.idf	17.066	$3.61e-05$	reject H_0
CREG-5K-OSU	stts semi-global+ tf.idf	14.2182	0.00016285	reject H_0

Table 6.14: Hypothesis test statistics for the baseline system augmented with STTS group part of speech and *tf.idf* weighting for cases of statistically significant system performance improvement.

system variant	3620-KU	3620-OSU	1032-KU	1032-OSU	5K-KU	5K-OSU
baseline	81.5	82.2	84.6	87.0	80.9	82.5
local + tf.idf	82.6	**84.4***	86.9	87.2	82.1	**84.5***
semi-global+ tf.idf	82.7	**84.3***	86.1	87.4	**82.8***	**84.2***
global+ tf.idf	**84.2***	**84.1***	85.1	85.8	**83.1***	**84.5***
ip+ tf.idf	82.4	**84.5***	84.9	86.5	**82.7***	**84.4***
lip+ tf.idf	83.0	**84.0***	85.4	84.8	**83.0***	**84.0***

Table 6.15: System performance for the baseline system augmented with part of speech and *tf.idf* weighting in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Question Types plus STTS Part of Speech Features plus tf.idf Weighting Table 6.17 shows the system performance of the short answer assessment system

Data set	feature variant	χ^2	p-value	decision
CREG-3620-KU	global+ tf.idf	5.4054	0.02007	reject H_0
CREG-3620-OSU	semi-global+ tf.idf	9.9448	0.001613	reject H_0
CREG-3620-OSU	local + tf.idf	11.9014	0.0005609	reject H_0
CREG-3620-OSU	lip + tf.idf	8.0958	0.004437	reject H_0
CREG-3620-OSU	ip + tf.idf	13.5619	0.0002308	reject H_0
CREG-3620-OSU	global+ tf.idf	8.2841	0.003999	reject H_0
CREG-5K-KU	global+ tf.idf	7.2727	0.007001	reject H_0
CREG-5K-KU	ip+ tf.idf	5.3645	0.02055	reject H_0
CREG-5K-KU	lip+ tf.idf	6.76	0.009322	reject H_0
CREG-5K-KU	semi-global+ tf.idf	5.083	0.02416	reject H_0
CREG-5K-OSU	global+ tf.idf	11.1692	0.0008317	reject H_0
CREG-5K-OSU	ip+ tf.idf	11.1304	0.0008492	reject H_0
CREG-5K-OSU	lip+ tf.idf	6.3018	0.01206	reject H_0
CREG-5K-OSU	local+ tf.idf	11.1122	0.0008576	reject H_0
CREG-5K-OSU	semi-global+ tf.idf	7.7173	0.005469	reject H_0

Table 6.16: Hypothesis test statistics for the baseline system augmented with part of speech and *tf.idf* weighting for cases of statistically significant system performance improvement.

augmented with all three weighting methods: question types, part of speech (STTS group) features, and *tf.idf* weighting. Table 6.18 shows the corresponding hypothesis test statistics. This weighting method combining all three levels of evidence has the best overall performance when compared to other feature set combinations. The feature set provides a total number of 17 statistically significant improvements across the data sets. For CREG-1032-KU the best result in all experiments was measured with an accuracy of 88.9% for a system with question type features, interpolated STTS part of speech features, and *tf.idf* weighting. In general, this feature set combination is one out of two feature sets only which can yield a statistically significant improvement on CREG-1032-KU. However, for CREG-1032-OSU, this feature set doesn't perform as well as other feature sets. The semi-global feature variants improve the performance to 88.2%, but apart from this improvement, the performance even drops with regard to the baseline. These 88.2% are also far away from the 90.0% reached when using part of speech features only.

system variant	3620-KU	3620-OSU	1032-KU	1032-OSU	5K-KU	5K-OSU
baseline	81.5	82.2	84.6	87.0	80.9	82.5
q-types + stts local + tf.idf	83.8	**84.7***	**87.9***	86.5	82.4	84.9
q-types + stts semi-global+ tf.idf	83.1	**84.6***	85.4	88.2	82.1	84.9
q-types + stts global+ tf.idf	**84.2***	**84.5***	**87.9***	84.6	**82.6***	**84.6***
q-types + stts ip+ tf.idf	83.3	**84.7***	**88.9***	84.1	**82.8***	**85.3***
q-types + stts lip+ tf.idf	**84.5***	**85.0***	**88.0***	85.8	**82.8***	**84.9***

Table 6.17: System performance for the baseline system augmented with question type and STTS group part of speech features and *tf.idf* weighting in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

Data set	feature variant	χ^2	p-value	decision
CREG-3620-KU	q-types + stts local + tf.idf	4.3478	0.03706	reject H_0
CREG-3620-KU	q-types + stts lip + tf.idf	5.6279	0.01768	reject H_0
CREG-3620-OSU	q-types + stts global + tf.idf	12.7735	0.0003516	reject H_0
CREG-3620-OSU	q-types + stts ip + tf.idf	14.4417	0.0001446	reject H_0
CREG-3620-OSU	q-types + stts lip + tf.idf	18.8876	$1.386e-05$	reject H_0
CREG-3620-OSU	q-types + stts semi-global + tf.idf	13.7357	0.0002104	reject H_0
CREG-3620-OSU	q-types + stts local + tf.idf	14.4868	0.0001411	reject H_0
CREG-1032-KU	q-types + stts global + tf.idf	5.5538	0.01844	reject H_0
CREG-1032-KU	q-types + stts ip + tf.idf	10.9649	0.0009285	reject H_0
CREG-1032-KU	q-types + stts lip + tf.idf	6.4516	0.01109	reject H_0
CREG-1032-KU	q-types + stts local + tf.idf	5.2319	0.02218	reject H_0
CREG-5K-KU	q-types + stts global + tf.idf	4.1965	0.04051	reject H_0
CREG-5K-KU	q-types + stts ip + tf.idf	5.3965	0.02018	reject H_0
CREG-5K-KU	q-types + stts lip + tf.idf	5.3493	0.02073	reject H_0
CREG-5K-OSU	q-types + stts global + tf.idf	4.1965	0.04051	reject H_0
CREG-5K-OSU	q-types + stts ip + tf.idf	5.3965	0.02018	reject H_0
CREG-5K-OSU	q-types + stts lip + tf.idf	5.3493	0.02073	reject H_0

Table 6.18: Hypothesis test statistics for the baseline system augmented with question type and STTS group part of speech features and *tf.idf* weighting for cases of statistically significant system performance improvement.

6.3.4 Analysis of Question Type Specific Part of Speech Alignment

In order to understand the interaction between question types and part of speech alignment preferences, an analysis of part of speech alignments given a certain question type was conducted. To account for effects of data sparsity, especially for questions with few instances, no distinction was made in the analysis between student and target side features. This analysis was conducted on the basis of the complete CREG-1032 data set. For every part of speech tag in the tag set, its semi-global alignment quantity was output. The part of speech tags were ranked according to their informativeness with WEKA (Hall et al., 2009). In this toolkit, the *AttributeSelection* filter was used to rank the features according to their information gain when used for predicting the binary diagnosis. Table 6.19 shows the ten most informative semi-global part of speech alignments per question type, together with the number of student answers per question.

Even with the limited data, some interesting patterns can be detected. For the *Yes/No* questions, words tagged as answer particles (*PTKANT*) are the most informative tags for the diagnosis when (not) aligned. For the *when* question type, the alignment of the *ADV* tag, among others assigned to temporal adverbs, is most informative. In case of the *who* question, the alignment of nouns (*NN*) is the most informative action. For cases where there are several questions, it is important to align nouns (*NN*,*NE*) with a certain characteristic (*ADJA*). For *how* questions, the alignment of a nominal entity (*NN*) of a certain quantity (*CARD*) and characteristic (*ADJA*) performing a certain action (*VVFIN*) is most informative. In case of *what* questions, which ask for a certain entity, the alignment of a noun (*NN*) is most informative.

This analysis exemplifies that the combination of question type features and alignment-based part of speech features not only improves the classification accuracy, but that this process also yields results interpretable from a linguistic point of view. The results however also underline the need to account for data sparsity effects via equivalence classes, as realized with the STTS and coarse part of speech groups.

q-type	instances	10 most informative Part of Speech tags
Alternative	7	VVPP, PPOSAT, PPER, PPOS, VMFIN, PRELAT, PIS, PIDAT, PIAT, PDS
How	144	NN, CARD, VVFIN, ADJA, ART, VAFIN, NE, PIAT, PRELS, PTKNEG
What	276	NN, KON, ADJA, VVPP, VVINF, APPRART, PIS, CARD, PTKNEG, PWAV
When	6	ADV, KOKOM, KOUS, NN, PIS, PWF, PIDAT, PWAV, PPOSAT, VAFIN
Where	9	PIDAT, PPER, PPOSAT, PRELAT, PIS, VVPP, PRF, PIAT, PAVDAT
Which	170	NN, ADV, VVPP, PTKNEG, VAFIN, NE, VAINF, CARD, KON, PIS
Why	174	NN, VVFIN, ART, APPR, PIAT, VAFIN, KON, NE, ADJA, KOKOM
Who	41	NN, VVINF, ADJD, VMFIN, PPER, PRELAT, PRELS, PPOS, PPOSAT, PTKANT
Yes/No	5	PTKANT, PPOSAT, PRELAT, PPOS, PIS, PPER, PIDAT, PRF, PIAT, PAV
Several	200	NN, NE, ADJA, PIAT, VMFIN, KON, PIS, VVPP, KON, PTKNEG

Table 6.19: Most informative part of speech alignments by question type.

6.3.5 Macro-Averages by Question Type

(Meurers, Ziai, Ott, & Kopp, 2011) reported macro-averages of the CoMiC-DE system for CREG-1032 obtained by grouping the predictions of the system by question type and computing the accuracy separately for every question type. In order to see whether the extensions of the system proposed in the present work could improve the accuracy for a certain question type, an analogous procedure was conducted.

Since section 6.3.3 showed that combining all three weighting methods yielded the best overall results, this feature set was chosen for the following exemplary analysis. The feature values for all 1032 instances from CREG-1032-KU and CREG-1032-OSU were combined and the final predictions were grouped by predicted question type. This explains the differences in the number of instances for each question. Table 6.20 shows the macro-averages for every question type and part of speech feature variant. The system used question type, *tf.idf*, and part of speech (STTS group) features. The last line also shows the micro average when no distinction is made between question types. Numbers printed in boldface indicate an improvement over the accuracy reported in (Meurers, Ziai, Ott, & Kopp, 2011).

q-type	**# instances**	**local**	**sglobal**	**global**	**ip**	**lip**
Alternative	7	0.57	0.57	0.57	0.57	0.57
How	144	**0.88**	**0.89**	**0.91**	**0.90**	**0.90**
What	276	**0.87**	**0.88**	**0.87**	0.85	**0.88**
When	6	**1.00**	0.83	**1.00**	0.83	0.83
Where	9	0.67	0.56	0.67	0.67	0.67
Which	170	0.91	0.92	**0.93**	0.92	0.92
Why	174	**0.84**	**0.84**	**0.84**	**0.83**	**0.84**
Who	41	**0.88**	**0.90**	**0.85**	**0.88**	**0.85**
Yes/No	5	0.80	0.80	0.80	0.80	0.80
Several	200	**0.86**	**0.83**	**0.83**	**0.86**	**0.85**
Micro	1032	**86.7**	**86.8**	**87.0**	**86.5**	**87.3**

Table 6.20: Macro-averages of the best system variant on CREG-1032 obtained by grouping results by question type.

As table 6.20 shows, the new features improved the performance for the cases where (Meurers, Ziai, Ott, & Kopp, 2011) reported major problems. They pointed out that

the *who* and *why* questions pose difficulties to the system due to the relatively high freedom of expression allowed for answers to questions of this type. Across all feature variants the system performance improved for these question types. Another question type that allows for a higher variation in the answer is the *how* question type. For this question type, the new system also outperformed the CoMiC-DE system by (Meurers, Ziai, Ott, & Kopp, 2011). The new approach also performs better across all instances for the *several* question type. Despite these clear improvements, not all problems could be solved. The problem that the givenness filter excludes material given in alternative questions from the alignment still leads to the lowest performance of the system for this question type. Also the performance decreased for the *where* question type. This may warrant the implementation of a named entity recognizer capable of recognizing and tagging locations for these cases.
This analysis showed that the new features helped the system overcome problems previously reported. This improvement can be attributed to a combination of question type features, part of speech features, and *tf.idf* weighting. The question type features encode a difference in alignment patterns with respect to other question types, the part of speech features encode classes of the aligned elements, and the *tf.idf* features encode the importance of the aligned elements. For example in the case of the *who* questions, the question type feature can indicate that nominal expressions should be aligned. The fulfillment of this constraint is encoded in the part of speech features. Since it is unlikely that the person asked for appears often in other texts, but often in the corresponding text, the *tf.idf* feature outputs a high weight, telling the system that it is important that this entity was aligned. The interaction of the new features has thus explanatory potential with respect to the increased performance.

7 Discussion and Related Work

The experimental results show that the new features presented in this work improve the accuracy of the CoMiC-DE system. The alternative research hypotheses stated at the beginning of section 6.1, section 6.2, and section 6.3 could thus be confirmed.
The experiments provided evidence that the combination of all weighting methods yielded the best accuracy and most improvements across all data sets. While the implementation of *tf.idf* features improved the system performance consistently, the combination of question types and part of speech features did not yield as good results as when combined with *tf.idf* weighting. However, the analysis of question type specific part of speech alignments revealed linguistically plausible differences across data sets. Part of speech features alone already push the boundary of fully automatic short answer assessment, resulting in the overall best accuracy of 90.0% on the CREG-1032-OSU data set. As will be shown in the remainder of this section, the obtained results are highly competitive with state-of-the-art approaches towards short answer assessment.
This work is part of the active research field of short answer assessment. (Ziai et al., 2012) and (Burrows et al., 2015) give extensive overviews about the field. While it is not feasible to discuss all the systems here, the following paragraphs will discuss the most similar or relevant approaches to short answer assessment from the perspective of this work. The interested reader may refer to the above named two articles for a complete overview of the field.
This work is based on the CoMiC-DE (Meurers, Ziai, Ott, & Kopp, 2011) system,

which was the first short answer assessment for German. Therefore, this system, which was presented in section 3, is the most closely related work.
CoMiC-DE originated from CoMiC-EN (Meurers, Ziai, Ott, & Bailey, 2011), with the language-specific models for English replaced by models for German. CoMiC-EN represented a re-implementation of the CAM system (S. M. Bailey, 2008), (S. Bailey & Meurers, 2008) in the UIMA (Ferrucci & Lally, 2004) framework. This system showed an accuracy of 88.4% on a test set of 255 student answers. The conceptual basis of both CoMiC variants is the Content Assessment Module (CAM, (S. M. Bailey, 2008), (S. Bailey & Meurers, 2008)). For a more detailed discussion of these systems, refer to section 3.
A first variant of the part of speech features used in the present work were first described in (Rudzewitz & Ziai, 2015). They implemented part of speech alignment weighting features in an adapted version of CoMiC-EN (Meurers, Ziai, Ott, & Bailey, 2011) for the related task of answer selection. While this work represents the conceptual foundation of the coarse part of speech features presented in this work, there are also certain differences. Firstly, the features in (Rudzewitz & Ziai, 2015) were implemented for English with a tag set for this language. Secondly, they used data from another domain (web text instead of university-level learner language). And thirdly, the system was used in another context, namely answer selection instead of short answer assessment.
(Horbach et al., 2013) present a short answer assessment system that can be seen as an attempt to re-implement the CoMiC-DE (Meurers, Ziai, Ott, & Kopp, 2011) system. They explore the impact of textual features on their system's performance. Although their goal is also the binary classification of student answers in CREG-1032, their work is only partly comparable to the present work. Firstly, in the re-implementation of the system, they make use of the Zurich parser (Sennrich, Schneider, Volk, & Warin, 2009) instead of the MaltParser (Nivre et al., 2007) for dependency parsing. Secondly, they actively make use of part of speech features in the alignment process by treating part of speech identity as one of the possible basic conditions for aligning elements in the student answers. Instead of using part of speech features to weight existing alignments as done in the present work, they use part of speech information to establish alignments. (Horbach et al., 2013) provide results when using a varying number of nearest neighbors in TiMBL (Daelemans et al., 2004). The data used for the classification is also not exactly the CREG-1032 corpus used in the present work, since they removed 143 "problematic" (Horbach et al., 2013, page 5) student answers and corresponding questions. They do not in any way explain the characteristic of "problematic" student answers discriminating them from "non-problematic" student answers. Also their baseline of 82.2% accuracy makes their work not fully comparable to the present work. The innovative characteristic of their approach is the usage of the reading comprehension text in the classification process. They explore the impact of using the sentence closest to the student and the target answer as a feature in the classification process. In the best case, their system reaches an accuracy of 84.4%.
The usage of the text in the classification process is also pursued in the present work. Instead of focusing on a single sentence in the text as done by (Horbach et al., 2013), the present work uses the complete reading comprehension text associated to a student answer in order to determine *tf.idf* weights for the aligned words in the student and target answer. This method proved to be effective, since already by only adding the text-based *tf.idf* weights, the system accuracy could be pushed to 86.1% (CREG-1032-

KU) and 88.4% (CREG-1032-OSU) (see table 6.5). Another advantage of the present work compared to (Horbach et al., 2013) is the fact that the system in the present work doesn't need any additional annotation as used by their system. (Horbach et al., 2013) employed three annotators that conducted the annotation of the relevant text sentences for every student and target answer. In contrast to that, the present work only makes use of already existing resources, thereby eliminating the need for expensive annotation.

(Hahn & Meurers, 2012) present the CoSeC-DE system, a short answer assessment system for German based on formal semantics. They implement the framework of Lexical Resource Semantics (LRS, (Richter & Sailer, 2003)) to abstract from the surface form to a formal semantic representation. This is done for both student and target answers, and the LRS representations are aligned. From these alignments, they derive information structural properties of the answers in order to overcome the problems of the basic givenness filter used in the CoMiC system (Meurers, Ziai, Ott, & Kopp, 2011). They show results for the classification accuracy on CREG-1032. The best result is an accuracy of 86.3% on the combined KU and OSU data set. As section 6.3.5 shows, the CoMiC-DE system with the new features outperforms the CoSeC-DE system on the CREG-1032 data set.

(Hou & Tsao, 2011) propose an automatic short answer assessment system that makes use of part of speech features and *tf.idf* measures in a Support Vector Machine-based approach. Their system extracts the part of speech tag of every preceding and following word in the student answer. A feature for every part of speech tag in preceding and following a word is used in their system. Furthermore they use the student answer as a document to compute a *tf.idf* measure in order to obtain weights representing the importance of words in the student answers. In combination with n-gram features they obtain a precision of 71.9% on their corpus consisting of 9 questions and 342 free-text answers. They however do not use learner language as data, but questions from a university course on automaton theory and formal languages.

(Leacock & Chodorow, 2003) discuss the system c-rater for short answer assessment. They make use of spelling correction, syntactic decomposition, pronoun resolution, stemming, extraction of morphological negation, and distributional similarity to obtain a canonical representation of a student answer that is then compared to such a canonical representation of a manually specified target answer. The process of comparing student answers to target answers is not conducted automatically via machine learning tools as in the present work, but hand-crafted rules are applied that compare aspects of the canonical representations of the answers.

(Mohler, Bunescu, & Mihalcea, 2011) present a short answer assessment system that is based on the alignment of dependency graphs. They extract dependency triples from the student and target answers after enhancing the dependency annotations output by the Stanford Dependency Parser (De Marneffe, MacCartney, Manning, et al., 2006) to include additional features, including the parts of speech of the elements in the dependency relation. The extracted dependency triples are represented as feature vectors encoding a total of 68 features. A score is computed for every pair of nodes in the student and target answer, and the dependency graphs are aligned. The alignment is used for the extraction of alignment-based features by applying different transformations in the alignment process. Similar to the approach of the CoMiC-DE system (Meurers, Ziai, Ott, & Kopp, 2011), (Mohler et al., 2011) extract alignment-based features. Among a range of features, they use the inverse document frequency *idf* to

align nodes. Similar to the present work, they compare the parts of speech of words in the answers, as well as the coarse part of speech classes of words in the student and target answer. Another parallel to the present work is that they exclude words given in the question from the alignment process. They however not exclusively rely on an alignment-based approach, but combine alignment similarity with lexical similarity measures, including a *tf.idf* weight. The final output of their system as produced by a support vector machine is not a binary diagnosis, but a value on a 5-point scale. The data used for the evaluation of their system is from the domain of computer science assignments for undergraduate students and consists of 2237 student answers to 80 questions. The data set is however not balanced, as it is the case for the CREG corpus (Ott et al., 2012). They report the results in the Pearson correlation metric and reach a correlation of 0.52 with human scores. Interestingly a feature analysis showed that the *tf.idf* feature decreased the performance of the system in terms of correlation, but increased the performance strongly when analyzed with regard to error rate.
(Nielsen, Ward, & Martin, 2008) use enhanced dependency parse structures as the basis for their short answer assessment system. They derive "facets" (Nielsen, Ward, & Martin, 2008, page 2) from the dependency parses of the student and target answers. These facets encode both dependency triples and thematic roles. Their data consists of 290 questions asking for science knowledge and 15,400 student responses. Various (unbalanced) subsets of this data are used for training and testing. One important difference between their data and the CREG corpus is that they corrected spelling errors in the process of transcribing the student answers. From the conceptual point of view their data partitioning procedure shows certain parallels. As done in the present work, they create a test set with unseen questions. They also exclude facets that represent information given in the question. This exclusion of facets stands in line with the exclusion of given material from the alignment as in the present work. They extract a variety of features from the facets. Among the syntactic features, they compare the part of speech tags of the head and dependent of the facet. They also use part of speech tags in the computation of an edit distance metric from the target to the student answer. The features are then evaluated by a decision tree classifier.
(Ziai & Meurers, 2014) annotated the CREG-1032 data set with the information structural notion of focus. Focus "identifies the part of a sentence addressing the current question under discussion in the discourse" (Ziai & Meurers, 2014, page 1) and is used to replace the simplistic givenness filter used in the CoMiC-DE system. They make use of the task context information represented in the CREG corpus (Ott et al., 2012) to perform the annotation. Among focus and answer type, they annotate the question surface form in this process. The question categories though slightly differ from the ones used in the present work, which were adapted from (Meurers, Ziai, Ott, & Kopp, 2011). They show that the focus annotation could be performed with substantial inter-annotator agreement and use the annotation to evaluate the performance of the CoMiC system when aware of the (manually annotated) focus of an answer given a question. They conducted experiments where they compared the results of CoMiC-DE on CREG-1032-OSU when making use of givenness, focus, and both. The best results were obtained when the system made use of both givenness and focus, resulting in an accuracy of 90.3% when using the annotation of the first human annotator, and resulting in an accuracy of 89.3% when using the annotation created by a second human annotator. These results represent the best classification performance to date reported for the CREG-1032-OSU corpus. One problem identified by (Ziai & Meurers,

2014) however is that in order to reach this accuracy, expensive human annotation was necessary, and that an automatic focus detection system would be needed for future work. The present work showed that by using only already existing information, a system can be created that performs comparably on this data set. As shown in table 6.1, the CoMiC-DE system when augmented with local (coarse group) part of speech features reached an accuracy of 90.0%. While the upper bound of 90.3% could not be reached, the CoMiC-DE system discussed in the present work outperformed the focus-aware CoMiC-DE system using annotations of the second human annotator. One important aspect of this discussion is that the present work reached this competitive results by only using existing information in contrast to using additional expensive human annotation.

8 Conclusion and Future Work

This work showed the extension of a short answer assessment with features weighting alignments by general linguistic and task-specific properties. Different feature variants and combinations of features were explored, experimentally tested, and validated by statistical significance testing. As an instance of a general linguistic weighting part of speech features were implemented. The implementation differed with respect to the equivalence classes of tags and with respect to the normalization method. A hybrid approach to combine general linguistic and task-specific properties in one feature was pursued with the *tf.idf* weighting. The implementation of question types represented a purely task-specific weighting.

The setup of the experiments was designed to be similar to previous studies ((Meurers, Ziai, Ott, & Kopp, 2011), (Ziai & Meurers, 2014)). Although the baseline was slightly higher (87.0% instead of 84.6% on CREG-1032-OSU), the new features nevertheless could improve the performance significantly.

The combination of all levels of evidence (part of speech, *tf.idf*, question types) yielded the best overall performance across data sets. This shows that the weighting methods interact in a useful way. The *tf.idf* weighting always could improve the performance by weighting aligned terms by their importance in the corresponding reading text. The implementation of question types alone and the combination of question type features and part of speech features could not yield as good results as when combined with *tf.idf* weighting. However, an analysis of question type specific part of speech alignments showed linguistically interpretable results. An analysis of macro-averages obtained by grouping results by question type showed that the new features helped the system overcome problems with question types that were found to pose higher difficulties due to a high variation in the expression of the answer.

The maximal accuracy of 90.0% could be reached on the CREG-1032-OSU data set with part of speech features as the only new components. Since related studies published results for this data set((Meurers, Ziai, Ott, & Kopp, 2011), (Horbach et al., 2013), (Hahn & Meurers, 2012), (Ziai & Meurers, 2014)), it could be shown that the proposed weighting yielded results highly competitive with state-of-the-art approaches, without requiring the costly human annotation performed in other studies.

An idea for future work is to determine equivalence classes not only by clustering across instances, but also clustering by the diagnosis of this instance. The data set can be split into correct and incorrect instances and part of speech tags can be clustered separately for the data set partitions, resulting in a detailed picture of alignment

patterns for correct and incorrect instances. Another approach to be tested in future work is the mapping of the part of speech tags to an universal tag set with implicitly fewer distinctions, eliminating the need for the creation of equivalence classes and the normalization.
The best results of the CoMiC-DE system with 90.3% on the CREG-1032-OSU data set were reported by (Ziai & Meurers, 2014) when the CoMiC-DE system used the information structural notion of focus. Future work has to test how well the focus-aware CoMiC-DE system performs when it not only uses the focus of an answer to narrow down the window for comparisons, but how well it performs when it uses the new features of the present work for the actual comparison of the relevant part of the answer.

9 Acknowledgments

I would like to thank Ramon Ziai for his constant and helpful support of my work in the SFB 833 A4 project. I would also like to thank Detmar Meurers for offering me the possibility to work in this research project. Finally I would like to thank Niels Ott for introducing me to the technical foundations of the discussed short answer assessment system.

References

Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 107–115).

Bailey, S. M. (2008). *Content Assessment in Intelligent Computer-Aided Language Learning: Meaning Error Diagnosis for English As A Second Language* (Unpublished doctoral dissertation). Citeseer.

Baldridge, J. (2005). The OpenNLP Project. *URL: http://opennlp. apache. org/index. html,(accessed 25 August 2015)*.

Brown, G. (1983). *Discourse Analysis*. Cambridge University Press.

Burrows, S., Gurevych, I., & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Daelemans, W., Zavrel, J., van der Sloot, K., & Van den Bosch, A. (2004). TiMBL: Tilburg Memory-Based Learner. *Tilburg University*.

Day, R. R., & Park, J.-s. (2005). Developing Reading Comprehension Questions. *Reading in a Foreign Language*, *17*(1), 60–73.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of lrec* (Vol. 6, pp. 449–454).

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923.

Ellis, R. (2009). Task-based language teaching: sorting out the misunderstandings. *International Journal of Applied Linguistics*, *19*(3), 221–246.

Ferrucci, D., & Lally, A. (2004). UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, *10*(3-4), 327–348.

Gale, D., & Shapley, L. S. (1962). College Admissions and the Stability of Marriage. *American Mathematical Monthly*, 9–15.

Götz, T., & Suhre, O. (2004). Design and implementation of the UIMA Common Analysis System. *IBM Systems Journal*, *43*(3), 476–489.

Gries, S. T. (2008). *Statistik für Sprachwissenschaftler*. Vandenhoeck & Ruprecht.

Gütl, C. (2007). e-Examiner: Towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems. In *Proceedings of the 2nd International Conference on Interactive Mobile and Computer Aided Learning* (pp. 1–10).

Hahn, M., & Meurers, D. (2012). Evaluating the Meaning of Answers to Reading Comprehension Questions A Semantics-Based Approach. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 326–336).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka Data Mining Software: An Update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Hamp, B., Feldweg, H., et al. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 9–15).

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., ... others (2014). Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *arXiv preprint arXiv:1403.0801*.

Horbach, A., Palmer, A., & Pinkal, M. (2013). Using the text to evaluate short answers for reading comprehension exercises. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)* (Vol. 1, pp. 286–295).

Hou, W.-J., & Tsao, J.-H. (2011). Automatic assessment of students' free-text answers with different levels. *International Journal on Artificial Intelligence Tools*, *20*(02), 327–347.

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, *37*(4), 389–405.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Madnani, N., Burstein, J., Sabatini, J., & O'Reilly, T. (2013). Automated Scoring of a Summary Writing Task Designed to Measure Reading Comprehension. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–168).

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

Meurers, D., Ott, N., Ziai, R., et al. (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. *Proceedings of Linguistic Evidence. Tübingen*, 214–217.

Meurers, D., Ziai, R., Ott, N., & Bailey, S. M. (2011). Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(4), 355–369.

Meurers, D., Ziai, R., Ott, N., & Kopp, J. (2011). Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment* (pp. 1–9).

Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 752–762).

Nielsen, R. D., Ward, W., & Martin, J. H. (2008). Learning to Assess Low-Level Conceptual Understanding. In *Flairs conference* (pp. 427–432).

Nielsen, R. D., Ward, W., & Martin, J. H. (2009). Recognizing Entailment in Intelligent Tutoring Systems. *Natural Language Engineering*, *15*(04), 479–501.

Nielsen, R. D., Ward, W., Martin, J. H., & Palmer, M. (2008). Annotating Students' Understanding of Science Concepts. In *Lrec*.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., ... Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, *13*(02), 95–135.

Ogren, P., & Bethard, S. (2009, June). Building Test Suites for UIMA Components. In *Proceedings of the workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (setqa-nlp 2009)* (pp. 1–4). Boulder, Colorado: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/W/W09/W09-1501`

Ott, N., Ziai, R., & Meurers, D. (2012). Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. *Multilingual Corpora and Multilingual Corpus Analysis*, *14*, 47.

Partee, B., Ter Meulen, A., & Wall, R. (2012). *Mathematical Methods in Linguistics* (Vol. 30). Springer Science & Business Media.

Richter, F., & Sailer, M. (2003). Basic Concepts of Lexical Resource Semantics. In *Esslli* (pp. 87–143).

Rudzewitz, B., & Ziai, R. (2015). CoMiC: Adapting a Short Answer Assessment System for Answer Selection. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval* (Vol. 15).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Schiller, A., Teufel, S., & Thielen, C. (1995). Guidelines für das Tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*, *66*.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing* (Vol. 12, pp. 44–49).

Sennrich, R., Schneider, G., Volk, M., & Warin, M. (2009). A New Hybrid Dependency Parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology*, 115–124.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., & Beck, K. (2012). Stylebook for the Tübingen Treebank of Written German (tüba-d/z).

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL.

Ziai, R., & Meurers, D. (2014). Focus Annotation in Reading Comprehension Data. *LAW VIII*, 159.

Ziai, R., Ott, N., & Meurers, D. (2012). Short Answer Assessment: Establishing Links Between Research Strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 190–200).